Thermal Time Shifting: Leveraging Phase Change Materials to Reduce Cooling Costs in Warehouse-Scale Computers

Matt Skach*, Manish Arora†‡, Chang-Hong Hsu*, Qi Li‡, Dean Tullsen‡, Lingjia Tang*, Jason Mars*

*University of Michigan – †Advanced Micro Devices, Inc. – ‡University of California, San Diego

skachm@umich.edu, manish.arora@amd.com, hsuch@umich.edu, liqi@eng.ucsd.edu, tullsen@eng.ucsd.edu, lingjia@eecs.umich.edu, profmars@umich.edu

Abstract

Datacenters, or warehouse scale computers, are rapidly increasing in size and power consumption. However, this growth comes at the cost of an increasing thermal load that must be removed to prevent overheating and server failure. In this paper, we propose to use phase changing materials (PCM) to shape the thermal load of a datacenter, absorbing and releasing heat when it is advantageous to do so. We present and validate a methodology to study the impact of PCM on a datacenter, and evaluate two important opportunities for cost savings. We find that in a datacenter with full cooling system subscription, PCM can reduce the necessary cooling system size by up to 12% without impacting peak throughput, or increase the number of servers by up to 14.6% without increasing the cooling load. In a thermally constrained setting, PCM can increase peak throughput up to 69% while delaying the onset of thermal limits by over 3 hours.

1. Introduction

Increasingly, a significant portion of the world's computation and storage is concentrated in the cloud, where it takes place in large datacenters, also referred to as "warehouse-scale computers" (WSCs) [1]. One implication of this centralization of the world's computing infrastructure is that these datacenters consume massive amounts of power and incur high capital and operating costs. Even small improvements in the architecture of these systems can result in huge cost savings and/or reductions in energy usage that are visible on a national level [1, 4, 14, 22, 23, 26].

Due to the increasing computing density of these systems, a significant portion of the initial capital expenditures and recurring operating expenditures are devoted to cooling. To prevent high server failure, the cooling infrastructure must be provisioned to handle the peak demand placed on the datacenter. The scale of cooling infrastructure can cost over 8 million dollars [14], even if the datacenter only reaches peak utilization for a fraction of a load cycle. The cooling system also may become inadequate as servers are upgraded or replaced and the thermal characteristics of the datacenter change.

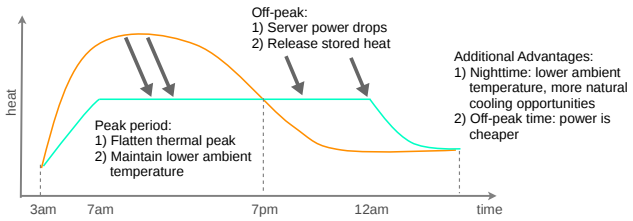

Figure 1: Thermal time shifting using PCM.

To mitigate these challenges, we propose the use of phase change materials (PCMs) to temporarily store the heat generated by the servers and other equipment during peak load, and release the heat when we have excess cooling capacity. The advantages of this approach may not be immediately obvious, because heat is not being eliminated, it is only stored temporarily then released at a later time. However, the key insight of this work is that the ability to store heat allows us to shape the thermal behavior of the datacenter, releasing the heat only when it is advantageous to do so.

This thermal time shifting is illustrated in Figure 1. This figure presents a diurnal pattern with a peak utilization and heat output during the middle of the day (7 AM to 7 PM). If we were able to cap heat output during the peak hours and time shift the energy until we have excess thermal capacity in the off hours, we can maintain the same level of server utilization using a cheaper cooling system with a much smaller cooling capacity.

This PCM-enabled thermal time shifting allows us to significantly reduce capital expenses, as we can now provision the cooling infrastructure for a significantly lower peak demand. Prior work on power shifting using batteries [8, 14] demonstrates the ability to produce a flat power demand in the face

ISCA'15, June 13-17, 2015, Portland, OR USA

of uneven diurnal power peaks. However, the power for the cooling still peaks with the workload. This work allows the cooling power to also be flattened, placing a tighter cap on total datacenter power.

Alternatively, we can use PCM to pack more computational capacity into the warehouse of an existing datacenter with a given cooling infrastructure without adding cooling capacity– this better amortizes the fixed infrastructure costs of the entire datacenter. Furthermore, given a load pattern such as the one in Figure 1, the ability to shift cooling demands from peak hours to the night time would allow us to take advantage of lower electricity rates during the night, or even leverage free cooling in regions with low ambient temperatures [3, 7, 8, 17, 37].

Despite the numerous advantages of PCM-enabled thermal time shifting, a number of important research challenges need to be addressed to fully exploit its advantages:

1. We need an adequate simulation methodology and infrastructure to study the PCM design space. To directly deploy PCM at a datacenter scale for design space exploration would be cost-prohibitive.
2. We need to investigate the trade-offs of various PCMs and identify the material that fits best in the datacenter environment. No prior work has studied PCM-enabled computation on this scale before, and selecting the correct PCM is critical to maximize impact while minimizing total cost of ownership (TCO).
3. We need to investigate suitable design strategies for integrating PCM in thousands of servers. Modern commodity servers are designed with excess cooling and interior space to allow for many applications, but there are ways to leverage this reconfigurability to enhance PCM performance.
4. We need to quantify the potential cost savings of using PCM. Datacenter cooling systems are very expensive, and even a small reduction can save hundreds of thousands or millions of dollars.

In this work, we present the advantages of PCM on a datacenter scale. We consider several PCMs for deployment in a datacenter, and select one for further investigation. We then perform a set of experiments with PCM on a real server, and validate a simulator with these tests. Using our validated simulator, we perform a scale out study of PCM on three different server configurations to predict the impact of PCM deployed in a datacenter. In an unconstrained datacenter, we find PCM enables a 12% reduction in peak cooling utilization or the deployment of 14.6% more servers under the same thermal budget. In a thermally constrained datacenter (e.g., more servers than the cooling system can cool), we find PCM can increase peak throughput by up to 69% while delaying the datacenter from reaching a thermal limit by over three hours.

The rest of the paper is organized as follows: Section 2 presents the integration of PCM in WSCs and the trade-offs of various PCMs; Section 3 presents our proposed PCM server model and its validation; Section 4 presents our test servers and methodology for the scale out study; and Section 5 presents our evaluation results. We consider related works in Section 6 and offer concluding thoughts in Section 7.

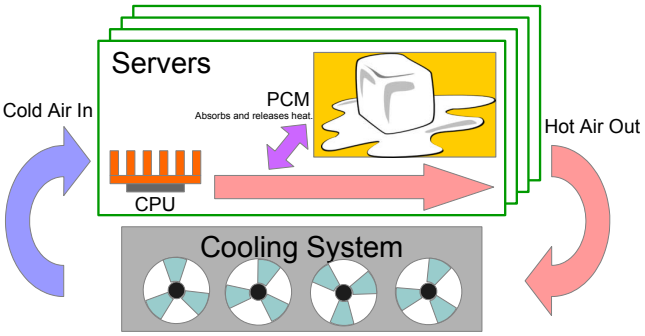

Figure 2: Integrating PCM in a WSC.

2. Integrating PCM in WSCs

To enable thermal time shifting, this work proposes to place a quantity of PCM inside of each server, as shown in Figure 2. When the temperature rises above the PCM's "melting threshold," the PCM will melt and absorb energy until all of the PCM is liquefied. Later, when the temperature drops below the threshold, the PCM will re-solidify and release energy until the PCM is solid again.

Placing PCM directly in contact with a the heat spreader of a single processor is beneficial for computational sprinting and other short-term cooling applications [29–31, 38], but we require a much greater quantity of PCM in a datacenter-sized cooling system with a 24 hour thermal cycle [13, 22]. Placing PCM in the server downwind of the processor sockets enables more PCM and still leverages the large temperature difference between idle and loaded levels. Alternatives such as placing PCM outside of the datacenter or adding a layer insulation in the walls and ceiling (reducing the ability of heat to escape when ambient conditions are favorable) require a infrastructure to move heat to the PCM and suffer a lower temperature differential due to heat loss and mixing over the travel distance.

Thus, the advantages of our PCM-enabled system are simple: the PCM is entirely passive. There is no power, software or floor space overhead to add PCM to a datacenter, and minimum labor is needed after installation to achieve the potential benefits.

2.1. Investigation of PCM Characteristics

A variety of PCM materials are available, but not all are suitable for the scale or operating conditions of a datacenter. To evaluate the available PCMs, several key properties need to be taken into account including the *melting temperature*, *energy density*, *stability*, and *cost*.

Melting temperature is critical as it determines when our PCM absorbs and releases significant amounts of heat. In a datacenter, we want the melting temperature to fall between

Table 1: Properties of common solid-liquid PCMs.

PCM	Melting Temp. (°C)	Heat of Fusion (J/g)	Density (g/ml)	PCM Stability	E. Conductivity	Corrosive?
Salt Hydates	25-70	240-250	1.5-2	Poor	High	Yes
Metal Alloys	>300	High	High	Poor	High	No
Fatty Acids	16-75	150-220	0.8-1	Unknown	Unknown	Yes
n-Paraffins	6-65	230-250	0.7-0.8	Excellent	Very Low	No
Commercial Paraffins	40-60	200	0.7-0.8	Very Good	Very Low	No

the peak and minimum load temperatures. Although the best melting temperature must be determined based upon ambient temperatures where the PCM is located, among other factors, the appropriate range is usually between 30 to 60 °C.

The energy density of the PCM defines how much energy it can store and is proportional to the heat of fusion (melting energy) and density of the PCM in both solid and liquid phases. A high energy density is desirable to maximize energy storage using the small amount of space available inside of the server. We also need to consider the corrosivity and electrical conductivity to contain a PCM and minimize damage in case it leaks out of the enclosure.

PCM Comparison - Of the phase transformations presented by Pielichowska, et al. [27], we find solid-liquid transformations to be promising for datacenter deployment right now. Liquid-gas and solid-gas have a much lower density in the gaseous state that reduces the energy storage density, and make PCM containment much more difficult. Solid-solid PCMs are attractive with a potentially high heat of fusion, low thermal expansion, and low risk of spillage; however, the solid-solid PCMs considered for energy storage by Pielichowska, et al. [27] undergo the phase change outside of acceptable datacenter temperatures, exhibit poor material stability in as few as 100 cycles of melting and resolidifying, possess a low energy density, or would be cost prohibitive in a datacenter at this time.

In Table 1 we compare five types of solid-liquid PCMs. Of the five, salt hydrates and metal alloys both have a high energy density but poor stability over repeated phase changes. The typical melting temperature of the metal alloys is much too high for datacenter use, and salt hydrates and fatty acids are both corrosive [11, 12, 27, 33].

We find that paraffin waxes are the most promising of the PCMs available right now. Paraffins typically have a low density but a good heat of fusion, are non-corrosive and don't conduct electricity. Paraffin is also highly stable, with negligible deviation from the initial heat of fusion after more than 1,000 melting cycles [27]. Paraffin wax is typically available in two forms: molecular pure n-paraffin (eicosane, tridecane, tetradecane, etc.) and commercial grade paraffin. Eicosane, previously studied for computational sprinting [30], has promising material properties including a high heat of fusion (247 J/g) and an appropriate melting temperature of 36.6 °C. However, we conclude that it is cost prohibitive to deploy at large volume in a datacenter. Sigma-Aldritch® quoted the mass production price of eicosane n-paraffin at $75,000 per ton. Even in a relatively small datacenter the cost of equipping every server with eicosane would be over a million dollars in wax costs alone.

Commercial grade paraffin is a less refined wax consisting of a mixture of paraffin molecules. It has a slightly lower heat of fusion (200 J/g), but is much less expensive than eicosane. As of August 2014, quotes for bulk commercial grade paraffin with melting temperatures ranging between 40 and 60 °C were typically $1,000 to $2,000 per ton on Alibaba.com® [24]: 50x cheaper for 20% lower energy per gram compared to eicosane, which we deem as a reasonable trade-off.

3. Modeling and Model Validation

The lack of experimental infrastructure and simulation methodology is a major challenge for conducting an investigation on PCM-enabled thermal time shifting. In this section, we introduce our infrastructure to simulate paraffin wax inside of a server. We integrate PCM modeling within a computational fluid dynamics (CFD) simulation for server layout using ANSYS® Icepak. To validate our PCM modeling, we rely on a series of measurements taken using a small quantity of paraffin inside of a real server and compare our model against those real server results. Modeling heat and airflow at this level is critical for two reasons. First, we need to accurately model heat exchange between the components, the air, and the wax. Second, our wax enclosures disrupt the airflow of the server and can have negative effect on heat removal if placed incorrectly.

Test System Configuration - We perform extensive benchmarking of a Lenovo® RD330 server to accurately model the server in Icepak and validate the model of PCM in Icepak. Our RD330 (Figure 3) is a 1U server with two sockets, each populated by a 6-core Intel® Sandy Bridge Xeon® CPU clocked at 2.4 GHz with Intel TurboBoost turned off. The server has 144 GB of RAM in 10 DDR3 DIMM sticks, a 1 TB 2.5" hard drive, and a single power supply unit rated at 80% efficiency idle and 90% efficiency under load. The server has six 17W fans, and runs Ubuntu® 12.04 LTS server edition. For the PCM, we purchased commercial grade Paraffin wax from Amazon.com® and measured the melting temperature at 39 °C.

Experimental Methodology - Accurate measurement is critical for creating an accurate model. To acquire accurate ground-truth measurement, we design several experiments and

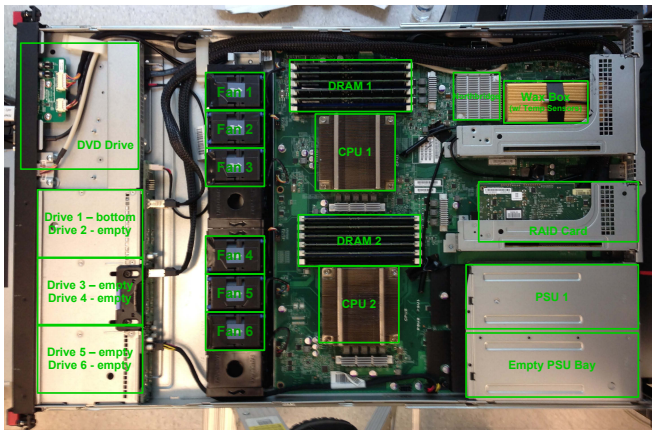

Figure 3: RD330 Server with major components labeled.

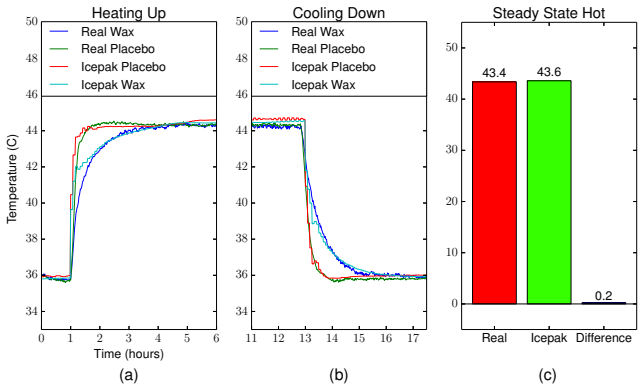

Figure 4: Model Validation. Transient traces while heating up (a) and cooling off (b), and steady state while hot (c) comparison of temperatures around the wax in the real server and our Icepak model.

use a number of tools to measure server power, temperatures at various points, and PCM's impact on temperatures. To measure total system power at the wall, we use a Watts Up? Pro® USB® power meter. We measure internal temperatures in the server with a set of TEMPer1 USB temperature sensors. We also use the Intel Power Governor tool to measure the socket, core, and DRAM power in real time.

To measure the effect of a small amount of PCM in the system, we fill a sealed aluminum container with 90 ml (70 grams) of paraffin wax and leave an extra 10 ml of airspace to account for paraffin expansion and contraction. The aluminum box was placed in the rear of the server, downwind of CPU 1 and three TEMPer1 sensors were inserted to record temperatures near the box and server outlet. We also conducted a series of trials with the same aluminum box empty of wax (filled only with air) in the same location in the server as a placebo to further validate our model as well as separate the thermal effects of the PCM and airflow impact of the box on the server.

We perform multiple trials with and without wax where we subject the server to 60 minutes of idle time, followed by 12 hours under heavy load (one instance of SPEC® h264 per logical thread) to heat the server up until temperatures stabilize, and then 12 hours at idle again to measure the server cooling down.

We observe that the total system power doubles from 90 W idle to 185 W fully loaded. CPU power increased by 7.7x from 6 W idle to 46 W per socket under load. Package temperature, as reported by the chip's internal sensors, rose from 42 °C idle to 76 °C under load.

Modeling Server and PCM in Icepak - To simulate the effects of wax in our server, we construct a model of our server in the computational fluid dynamic simulator ANSYS Icepak. From front to rear, we model the hard drive, DVD drive and front panel as a pair of block heat sources. The fans are modeled as a time-based step function between the idle and loaded speeds. Each DRAM module is modeled independently, but memory accesses are approximated as uniform to evenly distribute power across all of the modules. The PSU is modeled in the rear of the server enclosure, and all other heat sources (motherboard, LEDs, I/O, etc.) are lumped together with the CPU sockets.

Model Validation - In Figure 4 (a) and (b), we highlight the heating up and cooling down traces of average temperatures near the server outlet. We see a strong correlation between the real measurements and Icepak simulation measurements for the trace, and observe the wax reduces temperatures for two hours while the wax melts (absorbing heat), and afterwards increases temperatures for two hours while the wax freezes again (releasing heat).

In Figure 4 (c), we compare steady state temperatures measured from USB sensors on the real server to temperatures measured from the same locations on the Icepak model while both were fully loaded (between hours 6 and 12). We observe a mean difference of 0.22 °C between the real measurements and Icepak simulation measurements on the loaded server.

4. Methodology

In this section, we introduce our methodology and candidate machines for a scale out study on PCM datacenters. We examine three homogeneous datacenters each provisioned with a different type of machine, shown in Figure 5. First, we consider a deployment of low power servers using the same 1U commodity server validated in Section 3. Second, we consider a high-throughput deployment consisting of 2U commodity servers similar to the Sun® Server X4470 with four 8-core Intel Xeon CPUs, and last we consider a high-density deployment of Microsoft® Open Compute® blades with two 6-core Xeon CPUs each. We evaluate each datacenter using real workload traces from Google®, and present the results in Section 5.

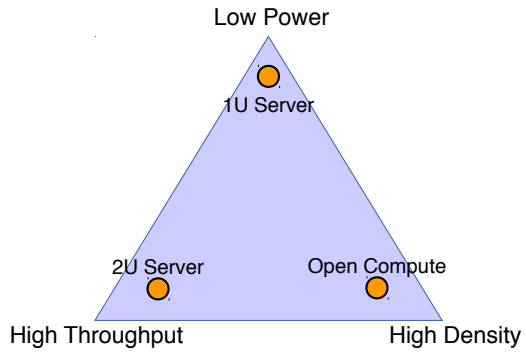

Figure 5: Three servers considered in the scale out study, each targeting a different end of the spectrum.

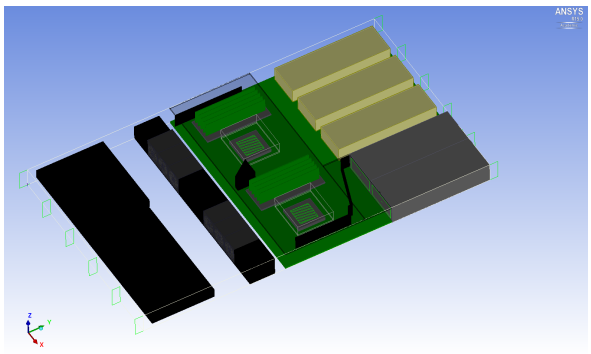

Figure 6: 1U low power server modeled in Icepak with 1.2 liters of wax (gold).

4.1. Servers

1U Commodity Server - The Lenovo RD330 we validated is a low power, 1U commodity server with an estimated cost of $2,000 for our configuration. To increase available space inside of the server, we replace the PCIe® risers and unnecessary RAID card (there is only one HDD in the server) with PCM. We conduct a series of experiments in Icepak blocking airflow with a uniform grille downwind of the CPU heat sinks, shown in Figure 7 (a). In these experiments, we maintain a constant frequency and power consumption to maintain parity across configurations. From 0% (no air blocked) up to 90% of air flow blocked, we observe a 14 °C increase in air temperatures at the outlet, and at no time do the CPU temperatures reach unsafe levels.

We model the addition of 1.2 liters of wax inside of aluminum boxes as shown in Figure 6 blocking 70% of airflow downwind of the CPUs. We could have increased the amount of wax (blocking further airflow), but found it was better to leave sufficient space between the boxes and edges of the server, thus maximizing surface area in contact with moving air in order to speed melting.

2U Commodity Server - The Sun X4470 is a high-throughput commodity server with up to four Intel E7-4800 processors. We model the server with four 8-core processors and 32 GB of RAM in two DDR3 DIMM packages per socket. In a 2U form factor we can fit up to 20 servers per rack and we estimate peak server power at 500 W per server after the PSU. Based on suggested retail prices, we estimate total cost to be $7,000 per server.

We model the 2U commodity server in Icepak in Figure 8. From front (left) to rear (right), air is pulled in through a series of fans, passes over the RAM, through the CPU heat sinks, past vacant PCIe card slots and out the rear of the server. The PCIe slots are present in the commodity server, but in our configuration they are not utilized so we leverage the free airspace to add wax into the server.

In Figure 7 (b) we plot temperature in the server as air is blocked by a uniform grille. When less than 50% of the air flow through our 2U commodity server is blocked we observe an almost negligible impact on outlet and CPU temperatures while at above 50% the temperature increases exponentially.

To add wax to our server without dangerously raising temperatures, we choose to add 4 one liter aluminum boxes filled with wax (colored gold in Figure 8) and maintain sufficient unfilled space to account for thermal expansion. These boxes block 69% of airflow through the server, increasing the outlet and CPU temperatures (with empty boxes) by less than 6 °C.

Open Compute Blade Server - The published production Microsoft Open Compute server is a 1U, sub-half-width blade with two sockets each containing a 6-core Intel Xeon processor and 64 GB of RAM in two DDR3 DIMM packages per socket. Two solid state drives (SSDs) connected via PCIe provide primary data storage, while four 3.5" 2 TB hard drives are present for redundancy. Each quarter-height Open Compute chassis fits 24 blades and has a total of six fans that draw air out the rear of the servers at less than 200 linear feet per minute at the rear of the blade. The peak power consumption for any single blade is limited to 300 W before the PSU, and the air temperature behind Socket 2 was measured at 68 °C. We model the idle power at be 100 W and active power at no more than 300 W. Based on current (August 2014) market trends we estimate cost per blade to be $4,000 [28].

We model the Open Compute server in Icepak based upon published dimensions and specifications for the form factor, CPUs, hard drives, and motherboard [16, 28, 34, 35], and estimate dimensions and power ratings for the SSDs based on the Fusion-io enterprise product line [5]. As with the commodity servers, additional heat sources in the Open Compute blade are lumped together with the CPUs. We do not model the volume or power requirements of the Catapult FPGA board [28].

In Figure 9, we present three Icepak models of the Open Compute configurations. Figure 9 (a) shows the production Open Compute configuration. We observe that even in a densely populated server like Open Compute, there is still useful space available where we can add wax without impacting airflow: along the sides of either CPU, plastic inserts

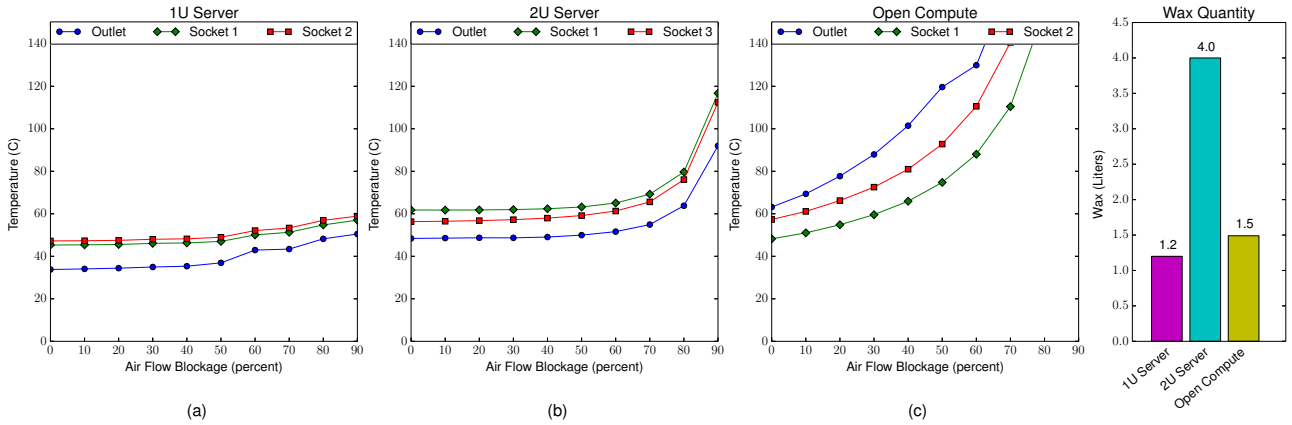

Figure 7: Server temperatures as airflow through each server is blocked. CPU temperatures in the 1U server (a) rise less than 2 °C below 50 %, and begin to rise quicker thereafter. Temperatures in the 2U server (b) are stable below 60 % quickly rise to unsafe rise to unsafe levels above 70 % obstructed airflow. Temperatures in the Open Compute server (c) rise to unsafe levels as soon as almost any airflow is obstructed.

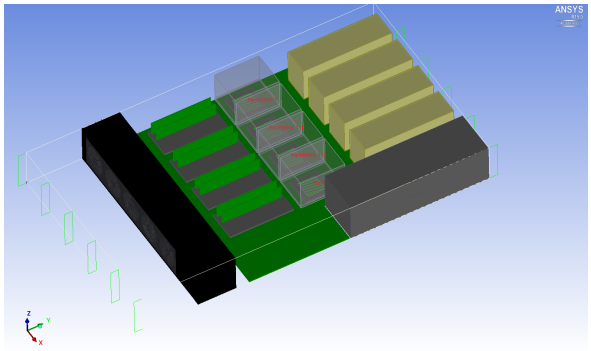

Figure 8: 2U high-throughput server with four CPU sockets modeled in Icepak with 4 liters of wax (gold).

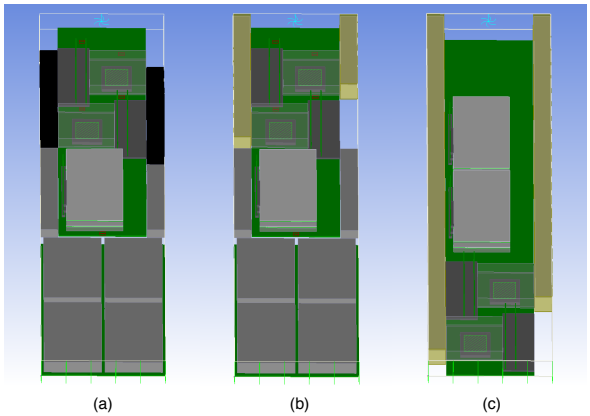

Figure 9: Icepak models of the Microsoft Open Compute server from [28] (a), Open Compute with air flow inhibitors replaced with wax containers (b), and Open Compute reconfigured with 1.5 liters of wax. (c).

(black) block air from traveling around the CPU heat sinks. In Figure 9 (b), we replace these blocks with 0.5 liters of wax in sealed aluminum containers.

The temperature gradient necessary to melt and cool wax in the server is created primarily by the CPUs, so wax is only useful if placed behind the CPUs. To increase the wax capacity, we consider an alternate configuration where we switch the CPU location with that of the SSDs to increase the downwind volume. We then consider a possible future Open Compute design where the redundant HDDs have been replaced with a second set of SSDs to achieve 1.5 liters of wax as shown in Figure 9 (c) without increasing the air flow blockage versus the production blade.

In Figure 7 (c), we study blocking additional airflow to add more than 1.5 liters of wax. (The outlet temperature is measured higher than CPU temperature due to the thermal output of the four enterprise class PCIe SSDs, which can exceed 85 °C even with proper cooling [40].) We observe that the already high outlet temperature and CPU temperatures increase exponentially as soon as any blockage is placed in the Open Compute blade, outweighing the benefits that any more wax would add.

4.2. Google Workload

We use a two day workload trace from Google [14, 36] to evaluate the effects of wax on our three datacenter server configurations. The workload we consider has three different job types: Web Search, Social Networking (Orkut®) and MapReduce from November 17th through November 18th, 2010. This data was acquired as described by Kontorinis, et al. [14], and normalized for a 50% average load and 95% peak load for a cluster of 1008 servers of each configuration. After 2011, Google changed the format of its transparency report so newer data is unavailable.

To model traffic and datacenter throughput, we use DCSim,

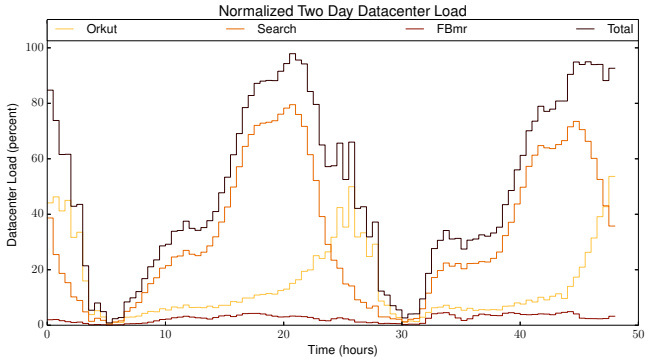

Figure 10: Two day datacenter workload trace from Google [14, 36] normalized to peak throughput.

a traffic-based simulator previously used by Kontorinis, et al. [14]. DCSim is an event-based simulator that models job arrival, load balancing, and work completion for the input job distribution traces at the server, rack, and cluster levels, then extrapolates the cluster model out for the whole datacenter. We use a round robin load balancing scheme, and extend DCSim to model thermal time shifting with PCM using wax melting characteristics derived from extensive Icepak simulations of each server.

4.3. TCO Modeling

We base our total cost of ownership (TCO) after Kontorinis, et al., modifying the model for our datacenter and server configurations, and add the interest calculation from Barroso, et al. (Table 2 and Equation 1) [1, 14]. To calculate the total savings from PCM, we consider the TCO without wax and subtract the TCO with wax for a single cluster of 1008 servers and extrapolate out to the size of the datacenter.

To best evaluate the TCO savings enabled by PCM, we consider the cooling infrastructure and the electricity cost of the cooling system separately from the datacenter operating expenditure (DatacenterOpEx). These two terms are important to our evaluation because they isolate the overall efficiency of the thermal-control system (including CRAC, cooling tower, and the PCM addition). We assume a linear relationship between the cost of cooling infrastructure and the peak cooling load the cooling system can handle. The electricity cost OpEx of the cooling system represents the average efficiency of removing heat. In addition, we also include the cost of adding the wax and the wax containers into the server capital expenditure (ServerCapEx), although the WaxCapEx is almost negligible representing less than 0.1% of the ServerCapEx.

To calculate the TCO for each server configuration, we consider three datacenters each with a critical power of 10 MW, the first filled with 55 clusters of 1U low power servers, the second with 19 clusters of 2U high throughput servers and the third with 29 clusters of Open Compute blades. We assume a peak electricity cost of $0.13 per kWh and an off-peak electricity cost of $0.08 per kWh [7].

Table 2: Parameters used to model TCO. (Dollars per watt refers to dollars per watt of datacenter critical power.)

Description	TCO/month	Unit
FacilitySpaceCapEx	1.29	$/sq. ft.
UPSCapEx	0.13	$/server
PowerInfraCapEx	15.9–16.2	$/kWatt
CoolingInfraCapEx	7.0	$/kWatt
RestCapEx	19.4–21.0	$/kWatt
DCInterest	31.8–36.3	$/kWatt
ServerCapEx	42–146	$/server
WaxCapEx	0.06–0.10	$/server
ServerInterest	11.00–38.50	$/server
DatacenterOpEx	20.7–20.9	$/kWatt
ServerEnergyOpEx	19.2–24.9	$/kWatt
ServerPowerOpEx	12.0	$/KWatt
CoolingEnergyOpEx	18.4	$/kWatt
RestOpEx	5.7–6.6	$/kWatt

5. Evaluation

In Section 3, we validated Icepak to simulate PCM in a server, and in Section 4, we described our servers and workload for a scale out study of PCM. In this section, we consider two potential use cases for PCM to reduce cooling load and increase throughput.

First, in Section 5.1 we consider a datacenter with a fully subscribed cooling system and evaluate how PCM can reduce the peak cooling load. This translates to a smaller, less costly cooling system or alternatively providing cooling support for more servers with the same cooling system. Next, in Section 5.2 we consider an oversubscribed datacenter and show how PCM can increase the datacenter throughput without surpassing the datacenter thermal threshold.

5.1. PCM to Reduce Cooling Load

We first consider a datacenter with a fully subscribed cooling system that can remove the peak cooling load indefinitely. The cooling load of a datacenter is the power that must be removed to maintain a constant temperature [2, 25], and allows a direct comparison between different server, temperature, and datacenter configurations. In Figure 11 (a-c), we plot the peak cluster cooling load for a cluster of 1008 of each test server without and with wax.

In this model, we assume all of the wax has a conservative heat of fusion of 200 J/g, and selected the melting temperature to minimize cooling load. The range of melting temperature available in commercial grade paraffin allows us to select one with an optimal melting threshold to reduce the peak cooling load of each cluster, and the best melting temperature is determined on the shape and length of the load trace: for

$$TCO = (FacilitySpaceCapEx + UPSCapEx + PowerInfraCapEx + CoolingInfraCapEx + RestCapEx) + DCInterest + (ServerCapEx + WaxCapEx) + ServerInterest + (DatacenterOpEx + ServerEnergyOpEx + ServerPowerOpEx + CoolingEnergyOpEx + RestOpex) \quad (1)$$

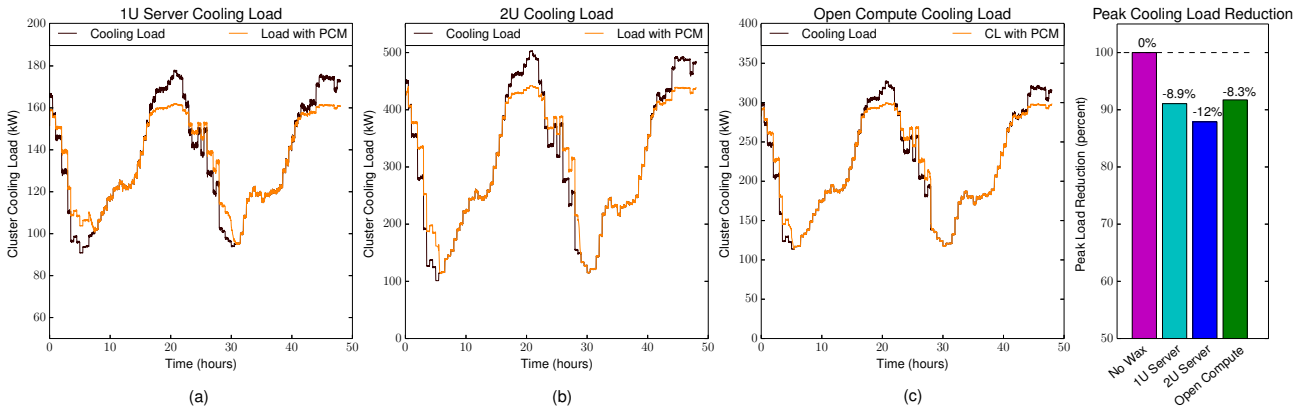

Figure 11: Cooling load per cluster over a two day Google trace in a datacenter with a fully subscribed cooling system. PCM reduces peak cooling load by 8.9 % in a cluster of low power 1U servers (a), 12 % in a cluster of 2U high throughput commodity servers (b), and by 8.3 % in a cluster of high density Open Compute servers (c).

the Google trace, we find that the best wax typically begins to melt when a server exceeds 75% load and melts quickly thereafter.

As shown, we achieve an 8.3% reduction in peak cooling in the Open Compute cluster, up to an 8.9% reduction in the cluster of 1U servers and 12% in the cluster of 2U servers as the wax absorbs heat and melts.

When the server utilization and temperatures fall below the melting threshold, we observe a period of time with increased cooling load higher than the placebo server while the wax cools off, lasting between six and nine hours. As the cooling system is operating below peak capacity during these times, there is sufficient cooling capacity to completely resolidify before the end of a 24 hour cycle.

With the peak cooling load safely reduced, we can then either decrease the size of the cooling system without sacrificing throughput, or add servers and increase critical power of the datacenter without increasing the size of the cooling system.

In a 10 MW datacenter, PCM allows us to install an 8.3% smaller cooling system in a high density Open Compute datacenter, an 8.9% smaller system with 1U low power servers, and a 12% smaller system with 2U high throughput servers. This translates to estimated cost savings of $174,000, $187,000, and $254,000 per year, respectively, on the cooling system and cooling power infrastructure. Here we observe that peak load reduction and savings correlate to the quantity of wax: the more wax that is added to a server, the greater the potential savings.

Alternatively, if instead of installing a smaller cooling system we use the excess cooling capacity enabled by PCM to install more servers, we can add 2,770 (8.9%) Open Compute blades, 4,940 (9.8%) more 1U low power servers or 2,920 (14.6%) more 2U high throughput servers to a 10 MW datacenter without exceeding the peak cooling load of the existing cooling system.

We evaluate the TCO savings created by oversubscribing the cooling system in a retrofit scenario: the old servers in a 10 MW datacenter have reached the end of their 4 year lifespan but the cooling system still has 6 years of useful lifespan remaining [14]. By adding PCM to a new deployment 1U, Open Compute, or 2U servers with an oversubscribed cooling system, we save an estimated $3.0 million, $3.1 million, and $3.2 million per year, respectively, over the cost of a new cooling system to achieve the same throughput.

5.2. PCM to Increase Throughput

In this section, we consider an oversubscribed datacenter where the cooling system is significantly smaller than the thermal output of the datacenter with all servers active. Such circumstances can arise as old servers are replaced with new denser servers, or in a datacenter constructed with an oversubscribed cooling system to run under peak power due to thread and cache contention issues, contention reducing techniques [15,21,39,42] that enable increased utilization through collocation increase the cooling load unsustainable.

In this oversubscribed datacenter, thermal management techniques such as downclocking/DVFS or relocating work to other datacenters [18–20] must be applied to prevent the datacenter from overheating.

In Figure 12, we plot the cluster throughput if the thermal limit did not exist and downclocking is not imposed, the

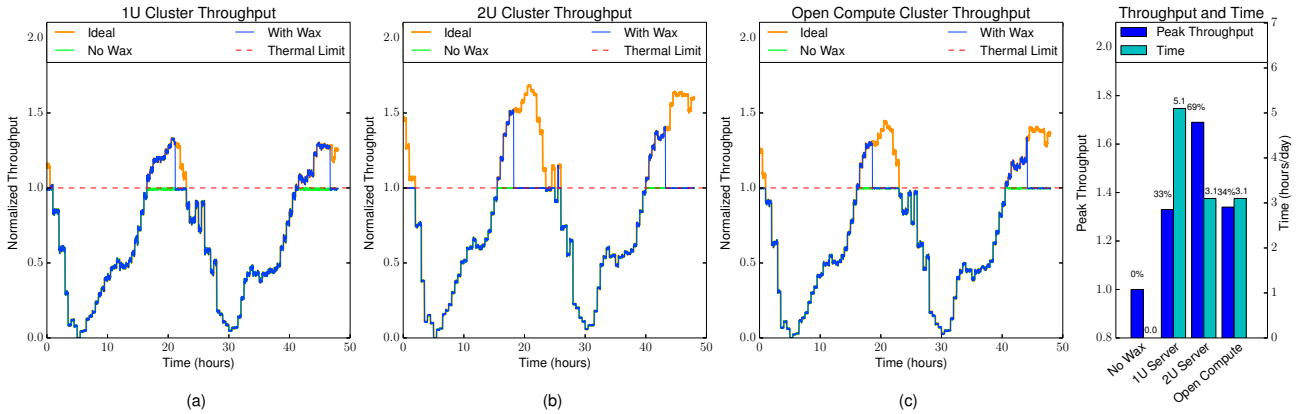

Figure 12: Google workload throughput normalized to peak throughput in a thermally constrained datacenter. PCM increases peak throughput by 33 % over 5.1 hours in the 1U server (a), 69 % over 3.1 hours in the 2U server (b) and 34 % over 3.1 hours in the Open Compute server (c).

throughput without wax, and the throughput with wax. In the trace without wax, downclocking to 1.6 GHz is imposed to prevent the cluster from overheating and throughput is normalized to the peak throughput while downclocked. Below the thermal limit, all three have the same throughput.

By adding PCM into the servers, we are able to maintain clock speeds and/or utilization as the wax absorbs thermal energy and until the thermal capacity of the wax is full. Once the wax is melted and can absorb no more energy downclocking or job relocation must be applied to prevent the datacenter from overhearing, but wax delays this by three to five hours.

In the Open Compute cluster, PCM delays the onset of thermal constraints by 3.1 hours and we observe a 34% increase in peak throughput during that time. In the 1U low power cluster, PCM delays thermal constraints by 5.1 hours with a 33% increase in peak throughput, and in the 2U high throughput cluster PCM delays thermal constraints by 3.1 hours and increases peak throughput by 69%.

To evaluate the impact of the increased throughput, we consider TCO efficiency: the ratio of TCO with increased peak throughput from PCM to the TCO required to achieve the same peak throughput without PCM. When thermal constraints lead to a decrease in throughput, we would need additional machines at significant additional cost to make up the difference. Thus an improvement without increasing the number of machines can lead to significant TCO efficiency savings.

We model TCO using Equation 1 with the assumption that most CapEx–including the facility space, power and the cooling infrastructure without PCM–are linear to the critical capacity of a datacenter [1]. OpEx terms related to the servers in Equation 1, such as server energy and cooling energy are proportional to the increase in the throughput and thus increase with or without wax.

In the 10 MW datacenter consisting of 1U low power servers, PCM achieves a TCO efficiency improvement of 23%, 39% in the 2U datacenter, and 24% in the high density Open Compute datacenter.

6. Related Work

The thermal energy storage potential of paraffin has previously been examined on a small, single-chip scale for computational sprinting in [29–31] with promising results. While that work uses PCM in small quantities to reshape the load without impacting thermals, we take the opposite approach, using PCM to reshape the thermal profile with minimal change to the load. Additionally, we study PCM deployment on a datacenter scale to consider thermal time shifting over periods lasting several hours, compared to seconds or fractions of seconds in the computational sprinting approach.

When considering PCM deployment across thousands of servers, we find that some of the techniques used in computational sprinting, such as the application of expensive n-paraffin wax, are cost prohibitive on our scale. We also observe that while Raghavan, et al. [30] studied a metal mesh embedded in paraffin to improve thermal conductivity, this potentially expensive measure is not necessary when melting paraffin over the course of several hours and the melting speed can be sufficiently improved by placing the paraffin in multiple containers to maximize surface area.

To reduce power infrastructure capital expenses in a datacenter, many authors have investigated UPS batteries to make up the difference when load exceeds the power distribution system power [8–10, 14, 37]. Our implementation of PCM is complementary to UPS power oversubscription.

Chilled water tanks for thermal energy storage is an active cooling solution considered by several authors [6, 32, 41, 43] to leverage the sensible heat of water during peak demand or emergencies. Our PCM approach is a completely passive thermal solution that is complementary to any active cooling solution (whether it be forced air HVAC, chilled water, etc.), because our passive technique will always reduce the peak demand placed on the active solution.

Comparing, in particular, to the chilled-water, active cooling solution of Zheng, et al. [43], PCM-enabled thermal time

shifting also has the advantage of no software, power or infrastructure overhead to control and contain water that TE-Shave requires. PCM requires no additional floor space or infrastructure because it is deployed inside of the server and draws no additional power, unlike chilled water tanks that must be deployed outdoors and cooled regularly, whether used or not, to compensate for environmental losses.

7. Conclusion

In this work, we introduce thermal time shifting, the ability to reshape a thermal load by storing and releasing energy when beneficial. We study paraffin wax, a phase change material that we place inside a real server to demonstrate thermal time shifting in a single server and validate a suite of software simulations we develop to study thermal time shifting on the cluster and datacenter scales. We show that thermal time shifting with a PCM can be used to reduce peak cooling load by up to 12% or increase the number of servers by up to 14.6% (5,300 additional servers) without increasing the cooling load. In a thermally constrained datacenter, we demonstrate that PCM can increase peak throughput by up to 69% while simultaneously postponing the onset of thermally mandated throughput reduction by over three hours.

Acknowledgements

The authors would like to thank the anonymous reviewers for many useful suggestions. This work was supported by NSF grants CCF- 1219059, CCF-1302682, and CNS-1321047. AMD, the AMD Arrow logo, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

References

[1] L. A. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 4, no. 1, pp. 1–108, 2009.

[2] A. Burdick, *Strategy Guideline: Accurate Heating and Cooling Load Calculations*. US Department of Energy, Energy Efficiency & Renewable Energy, Building Technologies Program, 2011.

[3] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2. ACM, 2007, pp. 13–23.

[4] M. E. Femal and V. W. Freeh, "Boosting data center performance through non-uniform power allocation," in *Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on*. IEEE, 2005, pp. 250–261.

[5] "Fusion-io products," http://www.fusionio.com/products/, 2014, Online; accessed 1-August-2014.

[6] D. Garday and J. Housley, "Thermal storage system provides emergency data center cooling," *White Paper Intel Information Technology, Intel Corporation*, 2007.

[7] Í. Goiri, W. Katsak, K. Le, T. D. Nguyen, and R. Bianchini, "Parasol and greenswitch: Managing datacenters powered by renewable energy," in *ACM SIGARCH Computer Architecture News*, vol. 41, no. 1. ACM, 2013, pp. 51–64.

[8] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and limitations of tapping into stored energy for datacenters," in *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*. IEEE, 2011, pp. 341–351.

[9] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters," in *ACM SIGPLAN Notices*, vol. 47, no. 4. ACM, 2012, pp. 75–86.

[10] ——, "Aggressive datacenter power provisioning with batteries," *ACM Transactions on Computer Systems (TOCS)*, vol. 31, no. 1, p. 2, 2013.

[11] D. Hale, M. Hoover, and M. ONeill, "Phase-change materials handbook," 1972.

[12] H. Ibrahim, A. Ilinca, and J. Perron, "Energy storage systems?characteristics and comparisons," *Renewable and Sustainable Energy Reviews*, vol. 12, no. 5, pp. 1221–1250, 2008.

[13] S. Kanev, K. Hazelwood, G.-Y. Wei, and D. Brooks, "Tradeoffs between power management and tail latency in warehouse-scale applications," *Power*, vol. 20, p. 40, 2014.

[14] V. Kontorinis, L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen, and T. Simunic Rosing, "Managing distributed ups energy for effective power capping in data centers," in *Computer Architecture (ISCA), 2012 39th Annual International Symposium on*. IEEE, 2012, pp. 488–499.

[15] M. A. Laurenzano, Y. Zhang, L. Tang, and J. Mars, "Protean code: Achieving near-free online code transformations for warehouse scale computers," in *Microarchitecture (MICRO), 2014 47th Annual IEEE/ACM International Symposium on*. IEEE, 2014, pp. 558–570.

[16] H. Li and A. Michael, "Intel motherboard hardware v2.0." Open Compute Project, 2011.

[17] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1. ACM, 2012, pp. 175–186.

[18] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 62–66, 2011.

[19] ——, "Greening geographical load balancing," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*. ACM, 2011, pp. 233–244.

[20] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation," *Performance Evaluation*, vol. 70, no. 10, pp. 770–791, 2013.

[21] J. Mars, L. Tang, R. Hundt, K. Skadron, and M. L. Soffa, "Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations," in *Proceedings of the 44th annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 2011, pp. 248–259.

[22] D. Meisner, C. M. Sadler, L. A. Barroso, W.-D. Weber, and T. F. Wenisch, "Power management of online data-intensive services," in *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*. IEEE, 2011, pp. 319–330.

[23] D. Meisner and T. F. Wenisch, "Peak power modeling for data center servers with switched-mode power supplies," in *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*. IEEE, 2010, pp. 319–324.

[24] "Paraffin wax listings on alibaba," http://www.alibaba.com/, online; accessed 22-July-2014.

[25] C. D. Patel, R. Sharma, C. E. Bash, and A. Beitelmal, "Thermal considerations in cooling large scale high compute density data centers," in *Thermal and Thermomechanical Phenomena in Electronic Systems, 2002. ITHERM 2002. The Eighth Intersociety Conference on*. IEEE, 2002, pp. 767–776.

[26] S. Pelley, D. Meisner, T. F. Wenisch, and J. W. VanGilder, "Understanding and abstracting total data center power," in *Workshop on Energy-Efficient Design*, 2009.

[27] K. Pielichowska and K. Pielichowska, "Phase change materials for thermal energy storage," in *Progress in Materials Science*, vol. 65, 2014, pp. 67–123.

[28] A. Putnam, A. M. Caulfield, E. S. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, G. P. Gopal, J. Gray *et al.*, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*. IEEE, 2014, pp. 13–24.

[29] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. Pipe, T. Wenisch, and M. Martin, "Utilizing dark silicon to save energy with computational sprinting," 2013.

[30] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin, "Computational sprinting on a hardwaresoftware testbed," vol. 48, no. 4. ACM, 2013, pp. 155–166.

[31] A. Raghavan, Y. Luo, A. Chandawalla, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin, "Computational sprinting," in *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*. IEEE, 2012, pp. 1–12.

[32] K. Roth, R. Zogg, and J. Brodrick, "Cool thermal energy storage," *ASHRAE journal*, vol. 48, no. 9, pp. 94–96, 2006.

[33] A. Sharma, V. Tyagi, C. Chen, and D. Buddhi, "Review on thermal energy storage with phase change materials and applications," vol. 13, no. 2. Elsevier, 2009, pp. 318–345.

[34] M. Shaw and M. Goldstein, "Open cloudserver blade specification v1.0." Open Compute Project, 2014.

[35] ——, "Open cloudserver chassis specification v1.0." Open Compute Project, 2014.

[36] "Google Transparency Report," http://www.google.com/transparencyreport/traffic/, 2011, Online; accessed 2011.

[37] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*. ACM, 2011, pp. 221–232.

[38] F. Volle, S. V. Garimella, and M. A. Juds, "Thermal management of a soft starter: transient thermal impedance model and performance enhancements using phase change materials," *Power Electronics, IEEE Transactions on*, vol. 25, no. 6, pp. 1395–1405, 2010.

[39] H. Yang, A. Breslow, J. Mars, and L. Tang, "Bubble-flux: Precise online qos management for increased utilization in warehouse scale computers," in *ACM SIGARCH Computer Architecture News*, vol. 41, no. 3. ACM, 2013, pp. 607–618.

[40] J. Zhang, M. Shihab, and M. Jung, "Power, energy and thermal considerations in ssd-based i/o acceleration," in *Advanced Computing Systems Association: HotStorage 2014*. USENIX, 2014.

[41] Y. Zhang, Y. Wang, and X. Wang, "Testore: exploiting thermal and energy storage to cut the electricity bill for datacenter cooling," in *Proceedings of the 8th International Conference on Network and Service Management*. International Federation for Information Processing, 2012, pp. 19–27.

[42] Y. Zhang, M. A. Laurenzano, J. Mars, and L. Tang, "Smite: Precise qos prediction on real-system smt processors to improve utilization in warehouse scale computers," in *Microarchitecture (MICRO), 2014 47th Annual IEEE/ACM International Symposium on*. IEEE, 2014, pp. 406–418.

[43] W. Zheng, K. Ma, and X. Wang, "Exploiting thermal energy storage to reduce data center capital and operating expenses," in *Proceedings of the IEEE 19th International Symposium on High-Performance Computer Architecture*. IEEE, 2014, pp. 132–141.